Modeling Accuracy and Coverage Evaluation Non-matches in the Census 2000

Michael Beaghen, Roxanne Feldpausch, Rosemary Byrne
Bureau of the Census
4700 Silver Hill Road
Suitland Federal Center, Bldg 2 Room 2407
Washington, D.C., 20233

KEY WORDS: logistic regression, E sample, P sample, erroneous enumeration

I. Introduction and Background[1]

The Accuracy and Coverage Evaluation (A.C.E.) measured the accuracy of the Census 2000 (Childers 2001). The A.C.E. consisted of two samples: a sample of people, the P sample, and a sample of census enumerations, the E sample. The P sample was obtained by conducting an independent enumeration of people in sampled clusters of census blocks. There were 721,734 P-sample people in 11,303 block clusters nationwide (Puerto Rico not included). The P sample measured the census miss rate. The E sample consisted of census enumerations in the A.C.E. sampled clusters. There were 712,900 E-sample people. The E sample measured the census erroneous enumeration rate. The miss rate and the correct enumeration rate, taken with the census count of enumerations eligible to be selected in the E sample, produced a dual system population estimate.

Central to the A.C.E. was a matching operation that compared the P-sample records to E-sample records and a field followup to resolve differences. People found in both the P sample and the census were called a match. People found only in the P sample and confirmed to be census day residents were called non-matches and represented census misses or failures to capture. E-sample people who matched to P-sample people who were residents were correct enumerations. E-sample people not matched to a person in the P-sample were followed up to determine whether they were correct enumerations or erroneous enumerations. If the non-matched person was found to have existed in the A.C.E. sample cluster on census day they were a correct enumeration. If the person was found not to have existed in the A.C.E. sample cluster on census day they were an erroneous enumeration. Examples of common erroneous enumerations were duplicated people and people not living as residents in the A.C.E. block cluster on census day. Also, census people with less data than a good name and two characteristics were counted aserroneously enumerated and were likewise not eligible to be matched to the A.C.E. The dual system estimates of population were unbiased as long as the census population to which the P-sample people could match was the population of census people meeting the A.C.E. definition of a correct enumeration.

The purpose of this paper was to build two logistic regression models: one to relate census misses as identified by P-sample non-matches to variables such as person demographic characteristics, housing unit characteristics and census enumeration methods, and a second to relate erroneous enumerations to similar variables. While univariate descriptive statistics were illuminating, they did not address the question of the relationship of one variable to the response in the context of other variables. A multivariate model avoided this limitation and thus complemented the univariate studies being done. Logistic regression was an appropriate multivariate method since the responses in both models were binary; e.g., for the P-sample model a person was either captured or missed by the census, for the E-sample model a person was a correct enumeration or an erroneous enumeration. The particular interest of this work was in gaining insights from multivariate analysis that were not seen in a univariate analysis.

A logistic regression model takes the following form; the response is defined as a success (i.e., census capture or correct enumeration) or failure (i.e., census miss or census erroneous enumeration). Logistic regression then models the natural logarithm of the odds of a success. The odds of success are related to the probability of a success by $p_i/(1 ! p_i)$ = odds, where p_i is defined as the probability of a success for the i^{th} individual. The k parameter logistic regression model is:

$$\log(p_i/(1 ! p_i))=\$_0+\$_1x_{i1}+þ+\$_kx_{ik}.$$

Note that this study was observational rather than experimental. The characteristics used as regressors in the model were not controlled by the researcher but rather were random variables in themselves. Consequently the modeling was not predictive but descriptive and causal

---

1 This paper reports the results of research and analysis undertaken by Census Bureau Staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties of ongoing research and to encourage disscussion of work in progress.

relationships between variables could not be inferred. Furthermore, the hypothesis tests used to determine which variables to include in the model were conditioned on the data and were therefore not strictly correct.

II. Study Methodology

We used SUDAAN's (RTI 1997) Proc Logistic to estimate the logistic regression models. SUDAAN correctly estimated models with data drawn from complex surveys such as the A.C.E., including calculating correct standard errors and test statistics. We also used SAS (1989) software's Proc Logistic to estimate the models. SAS gave the correct maximum likelihood estimates, that is, the same estimates that SUDAAN yielded, though the standard errors and test statistics SAS yielded were incorrect. However, SAS had more modeling capabilities than SUDAAN; for example, it generated useful statistics like the concordance rates. Also important, SAS was able to calculate estimates for models with large numbers of parameters, such as those with interaction terms. SUDAAN ran out of memory with such models and could not compute estimates. In SUDAAN we used the Taylor series method for variance estimation and the Wald chi-square statistic to test whether a variable was significant. We approached the modeling with backward selection. All variables turned out to be significant at the 0.1 alpha level in both the E-sample and P-sample modeling except for Form Type which was dropped from the modeling. The weighting for both the P-sample and E-sample models reflected the cluster sampling, large block subsampling and Targeted Extended Search sampling. The weighting for the P-sample model also reflected of the probability of residency and the non-interview adjustment (Ikeda 2000).

We investigated first order interactions. Because of the large number of interaction terms (29 choose two or 406 for the census modeling, and 30 choose two or 450 for the A.C.E. modeling), and SUDAAN's unsuitability for estimating models with many parameters, we screened for important interaction terms with estimates generated by SAS. The stronger interaction terms were then estimated and tested for significance using SUDAAN. While we were confident we found the most important first order interaction terms, we may have missed less important though significant ones. This limitation was not serious for this research because its purpose was descriptive not predictive. Further, with one exception the importance of interactions in the model was not large in any case.

The variables in the model included demographic and geographic ones such as the variables in the A.C.E. poststratification (Griffin 2000), in addition to housing unit variables. All variables were categorical. The variables used for both models are Tenure, Age-Sex, Racial-Ethnic Domain, Type of Enumeration Area and Region. For the E-sample we used the Number of Units at the Basic Street Address, which was classified into single unit, 2-9 units, or 10 or more units, consistent with the A.C.E. housing unit dual system estimation poststratification. For A.C.E. modeling we chose Type of Basic Address (TOBA) over FNHU, which was the number of housing units at a basic street address. These variables were nearly collinear with TOBA having slightly more explanatory power. For the E-sample model we also included though later dropped Form Type. We did not investigate A.C.E. operational variables because our interest was restricted to making inferences about the census.

For comparison to the multivariate results we generated univariate odds ratios. The univariate odds ratio for a variable was the odds ratio calculated with only that one variable in the model. For example, to calculate the univariate odds for Relationship, we estimated a logistic regression model with only the variable relationship as an independent variable.

III. Interpreting the Results of Modeling the E-Sample

Rather than examining the parameter weights themselves it was easier to interpret the odds ratios associated with an increase of one unit for each parameter, which were directly related to the parameter weights (Hosmer, Lemeshow 1989). Because the standard errors of the odds ratios are not symmetric about the estimate, 95% confidence intervals are shown instead. Since each of the variables had a value of zero or one depending on the level, the interpretation of the odds ratio is straightforward. The odds ratio for a variable refers to the ratio of the odds with the variable equal to one to the odds with the variable equal zero. The absolute value of the odds ratio only makes sense in comparison to reference levels. However, the odds ratios between levels of the same variable can be compared directly. As an illustration consider Relationship. As shown in Table 2, six of the Relationship categories were indicated by parameters which had estimates, while one category, reference person, was the reference level and had an odds ratio set to 1.000. Now, the odds ratio of 0.442 associated with Sibling implied that a person who was a sibling had only about 44% percent the odds of being correctly enumerated as the reference person, all other variables held constant. It also implied that a person

Table 1. Wald Chi Square Test Statistics for the E-Sample Model

Variables and Interactions	Wald Chi Square Test Statistic
Relationship	873.6
Size of Household	229.5
Units at Basic Street Address	185.8
Age-Sex	149.9
Tenure	102.8
Racial-Ethnic Domain	102.0
Region	30.6
Type of Enumeration Area	27.6
Age 18+, Relationship Child	40.6
Northeast, BSA 2-9 Units	27.0
Males 18-49, Black	3.5

who was a parent had about 28% greater odds (0.565/0.442) being correctly enumerated than a sibling had.

Table 2. Results of E-Sample Modeling

Variables	Model Odds Ratio	95% CI (Model)	Univariate Odds Ratio
Relationship			
Reference Person	1.000	reference level	1.000
Spouse	0.880	(0.848, 0.912)	1.302
Child	0.735	(0.658, 0.821)	1.008
Sibling	0.442	(0.399, 0.490)	0.463
Non-Relative	0.460	(0.433, 0.489)	0.491
Parent	0.565	(0.499, 0.639)	0.691
Other Relative	0.465	(0.447, 0.548)	0.679
Size of Household			
2 - 6 People	1.000	reference level	1.000
Single Person	0.626	(0.589, 0.666)	0.630
7 or more People	1.144	(1.015, 1.289)	0.908
Units at Basic Street Address			
Single Unit	1.000	reference level	1.000
2-9 Units	0.525	(0.478, 0.578)	0.365
10+ Units	0.645	(0.584, 0.713)	0.479
Age-Sex			
50+ Female	1.000	reference level	1.000
0-17	1.302	(1.156, 1.467)	1.100
18-29 Male	0.956	(0.883, 1.034)	0.609
18-29 Female	1.023	(1.022, 0.946)	0.688
30-49 Male	0.979	(0.918, 1.045)	1.133
30-49 Female	1.145	(1.077, 1.217)	0.933
50+ Male	0.834	(0.795, 0.875)	0.959
Tenure			
Owner	1.000	reference level	1.000
Non-Owner	0.706	(0.660, 0.755)	0.476
Racial-Ethnic Domain			
Non-Hispanic White	1.000	reference level	1.000
American Indian on Reservation	1.087	(0.897, 1.318)	0.964
American Indian off Reservation	0.773	(0.631, 0.947)	0.649
Hispanic	0.960	(0.889, 1.037)	0.725
Non-Hispanic Black	0.703	(0.654, 0.755)	0.542
Native Hawaiians or Pacific Islanders	0.775	(0.775, 0.555)	0.578
Asian	0.895	(0.796, 1.006)	0.752
Region			
West	1.000	reference level	1.000
Midwest	1.151	(1.050, 1.263)	1.262
Northeast	1.165	(1.045, 1.298)	0.957
South	0.918	(0.842, 1.001)	0.946
Type of Enumeration Area			
Large MSA, Mailout/Mailback	1.000	reference level	1.000
Medium MSA, Mailout/Mailback	1.030	(0.935, 1.133)	1.159
Small MSA & Non-MSA MO/MB	1.017	(0.914, 1.133)	1.197
All other TEAs	0.833	(0.748, 0.930)	1.128
Interactions			
Age 18+, Relationship Child	0.668	(0.590, 0.756)	N.A.
Northeast, BSA 2-9 Units	0.693	(0.604, 0.796)	N.A.
Males 18-49, Black	0.930	(0.861, 1.003)	N.A.

The Wald Chi-square test statistic (Table 1) gave an indication of the relative importance of a variable. Thus Relationship, with a Chi-square of 873.6 was a dominant variable.

Looking at Table 2, when we compared the model odds ratios with the univariate odds ratios we noticed that the

effects were sometimes stronger in the univariate analyses. (When comparing two odds ratios keep in mind that a larger value of the odds ratio indicated a stronger effect if both odds ratios were greater than one, but a weaker effect if both ratios were less than one). The odds ratio for Tenure in the model, 0.706, indicated a weaker effect than did the odds ratio estimated in univariate analysis, 0.476. For the Units at Basic Street Address variable, the odds ratios for the levels of multi-unit versus single unit given by the model (0.525 and 0.645) were also more modest than that estimated in the univariate model (0.365 and 0.479). Similar held true for the level non Hispanic Black of the variable Racial-Ethnic Domain (0.703 versus 0.542). This exaggeration of effect in the univariate model resulted because the univariate model did not take into account the correlations between regressor variables. In this case, Tenure, Units at Basic Street Address and Racial-Ethnic Domain were all correlated. Note that since interactions were not examined in univariate models there was no applicable (N.A.) univariate odds ratio for comparison for interaction terms.

There were two interactions worth pointing out. First, the variable Units at Basic Street Address showed that people in single units were more likely than people in multi-units to be correctly enumerated. The model odds ratios were 0.525 and 0.645 for people at basic street addresses with 2-9 units and 10 or more units. However, in the Northeast the odds of correct enumeration for people in structures with 2-9 units were even smaller as it was multiplied by a factor of 0.693. Also, adult children had even smaller odds (0.668) of correct enumeration than one would have predicted based on their age and relationship alone.

IV. Interpreting the Results of the P-sample Model

The interpretation of the P-sample model logistic regression results was analogous to that of the E-sample model. The variable Relationship stood out as an important variable (Table 3), with reference people and their spouses more likely to be captured by the census than other household members (Table 4). As in the E-sample model, the effects of several variables were shown by the P-sample model to be weaker than they appeared in the univariate analysis. These variables were Age-Sex, Tenure, Type of Basic Address, Racial Ethnic Domain, and Region. There were several interesting interaction terms. The most important was

Table 3. Wald Chi-Square Test Statistics P-Sample Modeling

Variables and Interactions	Wald Chi Square Test Statistic
Relationship	1521.5
Age-Sex	445.6
Size of Household	316.7
Tenure	277.1
Type of Basic Address	158.6
Racial-Ethnic Domain	110.0
Region	89.8
Type of Enumeration Area	47.6
Large Household, not Reference Person	154.2
Black or Hispanic, Midwest	33.6
Multi-unit, West	18.8

the one between the Size of Household and Relationship. People in large households had greater odds of capture, unless there were seven or more people, in which case people who were not the reference person had smaller odds of capture (0.457). This effect was due to the fact that the census form accommodated six people. Persons seven and higher were counted as roster people who were treated as not captured in the A.C.E. matching. Also, people in the Midwest were more likely to be captured than those in other regions, (1.472 vs. 1.093, 1.000 and 0.977), unless the person's ethnic group was Black or Hispanic, in which case they were about no more likely to be counted in the Midwest as in other regions ($1.472 \times 0.688 = 1.02$). Lastly, people in multi-units were better captured in the West than in other regions by a factor of 1.417.

V. Conclusions

Logistic regression was a useful method to examine what variables are associated with census misses and census erroneous enumerations. One gained insights one would not have in a univariate analysis.

This study provided some evidence as to efficacy of the A.C.E. poststratification. The poststratification was by Age-Sex, Racial Ethnic Domain, Tenure, Region, MSA/TEA, and Mail Return Rate. The model showed the predictive value of all these variables except Mail Return Rate, which was not included in our models. Noteworthy was that many of the variables associated with census capture were similarly associated with

Table 4. Results of P-Sample Modeling

Variables	Model Odds Ratio	95% CI (Model)	Univ-ariate Odds Ratio
Relationship			
Reference person	1.000	reference level	1.000
Spouse	1.063	(1.034, 1.092)	1.412
Child	0.827	(0.788, 0.868)	0.922
Sibling	0.532	(0.486, 0.581)	0.438
Parent	0.701	(0.639, 0.769)	0.816
Other relative	0.392	(0.366, 0.420)	0.345
Non-relative	0.423	(0.401, 0.446)	0.364
Missing	0.607	(0.537, 0.686)	0.605
Age-Sex			
50+ female	1.000	reference level	1.000
0-17	0.829	(0.771, 0.891)	0.625
18-29 Male	0.585	(0.548, 0.623)	0.477
18-29 Female	0.681	(0.638, 0.727)	0.391
30-49 Male	0.687	(0.650, 0.725)	0.807
30-49 Female	0.815	(0.771, 0.861)	0.638
50+ Male	0.823	(0.787, 0.861)	0.897
Size of Household			
2 - 6 People	1.000	reference level	2.630
Single Person	0.629	(0.597, 0.662)	1.761
7 or more People	1.103	(0.942, 1.292)	1.000
Tenure			
Owner	1.000	reference level	1.000
Renter	0.617	(0.583, 0.653)	0.431
Type of Basic Address			
Single	1.000	reference level	1.000
Multi-unit	0.671	(0.617, 0.729)	0.450
Trailer not in park	0.601	(0.516, 0.699)	0.475
Trailer in park	0.473	(0.375, 0.597)	0.413
Racial-Ethnic Domain			
White	1.000	reference level	1.000
American Indian on reservation	0.770	(0.618, 0.959)	0.536
American Indian off reservation	0.753	(0.611, 0.927)	0.443
Hispanic	0.828	(0.767, 0.894)	0.518
Black	0.698	(0.650, 0.750)	0.488
Native Hawaiian or Pacific Islander	0.566	(0.383, 0.837)	0.399
Asian	0.872	(0.771, 0.987)	0.708
Region			
West	1.000	reference level	1.000
Northeast	1.093	(0.974, 1.225)	1.041
Midwest	1.472	(1.315, 1.649)	1.468
South	0.977	(0.882, 1.082)	0.948
Type of Enumeration Area			
Large MSA, Mailout/Mailback	1.000	reference level	1.000
Medium MSA, Mailout/Mailback	1.091	(1.001, 1.184)	1.248
Small MSA & Non-MSA MO/MB	1.010	(0.924, 1.104)	1.248
All other types of enumeration areas	0.735	(0.671, 0.804)	0.973
Interactions			
Large Household, not Reference Person	0.457	(0.404, 0.517)	N.A.
Black or Hispanic, Midwest	0.688	(0.606, 0.781)	N.A.
Multi-unit, West	1.417	(1.211, 1.660)	N.A.

correct enumeration. These variables were relationship; reference people and their spouses were both missed less and erroneously enumerated less; Tenure, owners were both missed less and erroneously enumerated less by the census; Racial-Ethnic Domain, whites were both missed less and erroneously enumerated less; Region, Midwesterners were both missed less and erroneously enumerated less; and Type of Enumeration area, people in mailout/mailback areas were more likely to be both missed less and erroneously enumerated less. On the other hand, the Age-Sex variable did not fit this pattern.

VI. References

Childers, Danny R. (2001): *Accuracy and Coverage Evaluation: The Design Document*. DSSD Census 2000 Dress Rehearsal Memorandum Series, Chapter S-DT-1.

Griffin, Richard, & Haines, Dawn (2000): *Accuracy and Coverage Evaluation Survey: Final Post-stratification Plan for Dual System Estimation.* DSSD Census 2000 Procedures and Operations Memorandum Series #Q-24.

Hosmer, David W., Lemeshow, Stanley (1989): *Applied Logistic Regression*. John Wiley & Sons, New York.

Ikeda, Michael (2000): *Accuracy and Coverage Evaluation Survey: Specifications for the Missing Data Procedures.* DSSD Census 2000 Procedures and Operations Memorandum Series, Chapter Q-25.

RTI (1997): *SUDAAN User's Manual, Release 7.5.* Research Triangle Institute, Research Triangle Park, North Carolina, 27709.

SAS (1989): *SAT/STAT User's Guide, Version 6, Fourth Edition Volume 2*. The SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513.

Wolter, Kirk M. (1986): *Some Coverage Error Models for Census Data*. Journal of the American Statistical Association, June 1986, Vol. 81, No. 394.